DOCUMENT RESUME

ED 344 935 TM 018 297

AUTHOR Davison, Mark L.; Chang, Yu-Wen
TITLE Measurement Accuracy: An Application of Multidimensional Item Response Theory to the Woodcock-Johnson Psycho-Educational Battery-Revised Achievement Scales.
PUB DATE Apr 92
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Adults; Equations (Mathematics); Error of Measurement; Estimation (Mathematics); Information Theory; *Item Response Theory; *Mathematical Models; Mathematics Tests; *Measurement Techniques; Multidimensional Scaling; Reading Tests; Scores; Test Validity; Vectors (Mathematics)
IDENTIFIERS *Ability Estimates; Accuracy; *Woodcock Johnson Psycho Educational Battery

ABSTRACT

A two-dimensional, compensatory item response model and a unidimensional model were fitted to the reading and mathematics items in the Woodcock-Johnson Psycho-Educational Battery-Revised for a sample of 1,000 adults aged 20-39 years. Multidimensional information theory predicts that if the unidimensional abilities can be represented as vectors in the two-dimensional solution, then the multidimensional model can be used to obtain ability scores with smaller standard errors. In reading, the multidimensional model yielded scores with smaller standard errors, but multidimensional scores from subtests within reading were identical to the overall reading score. In mathematics, unidimensional scores were non-linearly related to multidimensional ability estimates, and for some subjects, multidimensional ability scores had larger standard errors. There is a 16-item list of references. One table and 12 figures provide analysis information. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARK L. DAVISON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED344935

Measurement Accuracy: An Application of Multidimensional Item Response Theory to the Woodcock-Johnson Psycho-educational Battery-Revised Achievement Scales

Mark L. Davison and Yu-Wen Chang

University of Minnesota

Paper to be presented to the American Educational Research Association, April 20, 1992, San Francisco CA.

Requests for reprints should be addressed to Mark L. Davison, Department of Educational Psychology, University of Minnesota, 178 Pillsbury Dr. S.E., Minneapolis, MN 55455.

TM018297

BEST COPY AVAILABLE

Abstract

A two-dimensional, compensatory item response model and a unidimensional model were fitted to the Reading and Mathematics items in the Woodcock-Johnson Psycho-educational Battery-Revised for a sample of 1000 adults. Multidimensional information theory predicts that, if the unidimensional abilities can be represented as vectors in the two-dimensional solution, then the multidimensional model can be used to obtain ability scores with smaller standard errors. In Reading, the multidimensional model yielded scores with smaller standard errors, but multidimensional scores from subtests within reading were identical to the overall Reading score. In Math, unidimensional scores were nonlinearly related to multidimensional ability estimates, and for some subjects, multidimensional ability scores had larger standard errors.

Measurement Accuracy: An Application of Multidimensional Item Response Theory to the Woodcock-Johnson Psycho-educational Battery-Revised Achievement Scales

Most applications of item response theory are based on unidimensional models. Most item pools are multidimensional. To deal with multidimensionality, items can be divided into unidimensional (or essentially unidimensional, Stout, 1990) subsets, and each subset can be analyzed separately using a unidimensional analysis. Luecht and Miller (in press) argue for this approach over a truly multidimensional analysis on three grounds: unidimensional parameter estimates are more stable than multidimensional estimates; unidimensional ability estimates are more interpretable; and unidimensional models will fit unidimensional subsets of items as well as a multidimensional model.[1]

When items measuring different abilities are positively intercorrelated, and unidimensional abilities can be represented as linear composites of multidimensional ones, then multidimensional information theory (Reckase, 1986) suggests that all of the items in the pool will contribute information about a given ability, not just the items designed to measure the given ability. For instance, if items in a battery measuring reading and math are all positively intercorrelated, then both the math and reading items contribute information about math ability; both

reading and math items contribute information about reading ability. Ackerman and Davey (1991) used the term collateral information to describe the information provided by items measuring one domain--say reading--to estimates of ability in a different domain--say math. Hereafter, we will distinguish between secondary items, those which contribute collateral information, from primary items which were specifically designed to measure the ability in question. In achievement batteries, collateral information is potentially most important in very specialized areas--say measurement or algebra--for which there are relatively few primary items, but many correlated secondary items.

In multidimensional ability estimation, both primary and secondary items are used, and hence collateral information is captured. Luecht and Miller (in press) have cautioned that, in practice, the MIRT models may not deliver the theoretically predicted advantages of collateral information. First, multidimensional parameters can be estimated less accurately than unidimensional parameters. Second, after items have been divided into unidimensional subsets, responses to items in any subset can be predicted about as precisely from either a unidimensional or a multidimensional model. Third, in their opinion, their unidimensional estimates seemed to better represent the intercorrelation of their two simulated abilities, although in our opinion this problem might be handled by a more optimal

rotation of their multidimensional solution. Finally, they observed a regression to the mean bias in their multidimensional estimates. This bias may be inherent in multidimensional estimates, but it probably is specific to their particular estimation scheme. Their study did not directly address the question of whether MIRT models can yield more accurate ability estimates by capturing collateral information.

In an attempt to avoid the pitfalls and capture the benefits of MIRT models, Ackerman and Davey (1991) investigated a unidimensional approach to capturing collateral information in adaptive testing. For each of two simulated dimensions, they performed a unidimensional calibration which included both primary and secondary items. They then applied a unidimensional estimation of ability, once using just primary items and once using both primary and secondary items. Their results illustrate the theoretical proposition (Lord, 1983) that both bias and the standard error of ability estimates should decrease with the addition of secondary items to capture collateral information. The correlation of estimated abilities increased with the addition of secondary items. The authors called for more research to determine whether and under what conditions the theoretically predicted gains in measurement accuracy could be achieved in practice.

Davey and Hirsch (1991) performed a simulation study to assess whether including both primary and secondary items in

unidimensional ability estimation would improve the accuracy of those estimates. As compared to including just primary items, including both primary and secondary items led to smaller standard errors of measurement, but an increased regression toward the mean bias. It is unclear whether this bias is inherent in the use of collateral information or whether it is specific to their particular estimation scheme (their method of setting the unit and origin of the multidimensional scores).

Whereas Ackerman and Davey (1991) and Davey and Hirsch (1991) have attempted to capture collateral information with unidimensional models, we have used a multidimensional approach. This paper focuses on the accuracy of ability estimates and only secondarily on issues such as parameter estimation and fit. Particularly, this paper focuses on how the number of items, the intercorrelation of items, and content features of the test seem to influence theoretically expected gains in information through capture of collateral information. Like the earlier work by Ackerman (1991), it suggests that estimated unidimensional abilities are not always simple linear combinations of multidimensional ones; hence, in practice, there can be a violation of the assumption underlying the argument that multidimensional approaches must yield more information.

Unidimensional and Multidimensional Information

Like the earlier work of Ackerman & Davey (1991) and Davey

and Hirsch (1991), our work is based on the multidimensional, two parameter logistic (M2PL) model,

$$P(\theta_j) = P(x_{ij} | a_i, d_i, \theta_j) = \frac{\exp(a_i \theta_j + d_i)}{1 + \exp(a_i \theta_j + d_i)} \quad (1)$$

where x_{ij} is the score (0, 1) on item i by person j, $P(\theta_j)$ is the probability of a correct response to item i by person j, a_i is the vector of item discrimination parameters, d_i is a scalar parameter related to the difficulty of the item, and θ_j is the vector of ability parameters for person j. If the vectors a_i and θ_j are of dimension one, and hence scalars, this model reduces to the standard unidimensional 2-parameter logistic model. See Reckase (1985, 1986) for an interpretation of the multidimensional parameters.

Reckase (1986) proposed a definition of multidimensional information (MINF) which is a direct generalization of the unidimensional item response theory (UIRT) concept. For UIRT, information at an ability level, θ, is defined as the ratio of the square of the slope of the item response function (IRF) to the variance of the error of measurement at θ:

$$I_i(\theta) = \frac{[\delta P_i(\theta)/\delta(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (2)$$

where $I_i(\theta)$ is the information from item i given ability θ, $P_i(\theta)$ is the probability of a correct response to item i for a person with ability θ, and $Q_i(\theta) = 1 - P_i(\theta)$.

In MINF, an ability corresponds to a direction in a space defined by k ($k = 1, \ldots, M$) reference vectors. MINF is defined by the mathematical function of Equation 2, but a directional derivative is substituted for the standard derivative in the numerator. For any ability, the directional derivative for the M2PL model is given by

$$\begin{aligned} \nabla_{\alpha}' &= \Sigma_k a_{i\,k} P_i(\theta) Q_i(\theta) \cos\alpha_k \\ &= P_i(\theta) Q_i(\theta) \Sigma_k a_{i\,k} \cos\alpha_k \ , \end{aligned} \qquad (3)$$

where α is the vector of angles between the reference vectors and the vector corresponding to the ability of interest, α_k is an element of that vector expressing the angle between reference vector k and the specified ability vector, and $a_{i\,k}$ is the discrimination parameter for item i on reference vector k. Substituting the directional derivative in Equation 3 for the derivative in Equation 2 yields

$$\underline{I}_{i\alpha} = \frac{[\underline{P}_i(\theta)\underline{Q}_i(\theta)\Sigma_k \underline{a}_{ik}\cos\alpha_{ik}]^2}{\underline{P}_i(\theta)\underline{Q}_i(\theta)} \qquad (4)$$

$$= \underline{P}_i(\theta)\underline{Q}_i(\theta)[\Sigma_k \underline{a}_{ik}\cos\alpha_i]^2$$

where $\underline{I}_{i\alpha}$ is the information of item $\underline{i}$ in direction α.
For the specified ability α, the standard error of measurement is

$$\underline{SE}_\alpha(\underline{\theta}) = \frac{1}{\sqrt{\underline{I}_\alpha(\theta)}} \qquad (5)$$

where $\underline{I}_\alpha$ is the total test battery information in ability direction α. $\underline{I}_\alpha$ is the sum of the item information in ability direction α:

$$\underline{I}_\alpha = \Sigma_i \underline{I}_{i\alpha} \quad . \qquad (6)$$

In traditional unidimensional IRT, only items measuring a specific domain, the primary items, are used to estimate a given ability. Hence, the sum in Equation 6 runs only over primary items in the test battery. In a truly multidimensional application, all items in a test battery, not just the primary items, are used to estimate any given ability. Hence, the sum

runs over primary and secondary items; the "total test information" includes information from primary items, but also collateral information from secondary items. Assuming the collateral information, the sum of the information from secondary items, is nonzero then multidimensional information must exceed unidimensional information. As Luecht and Miller (in press) correctly caution, this gain in information will be obtained only if the multidimensional parameters are reasonably well estimated. Common sense would suggest that the gains in information from a multidimensional approach would be most substantial (a) when the number of primary items is small so that the ability cannot be estimated precisely from those items alone, (b) when the number of secondary items is large, and (c) when the intercorrelations of primary and secondary items are large.

Like factor analytic solutions, multidimensional estimates of item discrimination and person ability parameters are determined only up to a rotation and change of scale. This indeterminacy leads to two important points. First, the discrimination and ability estimates provided by MIRT programs correspond to reference vectors in a multidimensional space, but they do not necessarily correspond to the most interpretable directions in the space, and hence to the abilities of interest. Second, in practice, one must find the most interpretable directions in the space (i.e. the directions corresponding to the abilities of interest).

Existing MIRT programs would not provide estimates of math ability. Rather they would provide estimates of discrimination parameters and person abilities along reference vectors, from which one can hopefully estimate the angle between the reference vectors and the math ability vector and then estimate person projections onto the vector corresponding to math ability. Implied in the previous sentence is an important assumption: that unidimensional abilities are linear functions of multidimensional reference vectors. Hence, a MIRT application includes the following three steps; estimating item discrimination and person abilities along reference vectors, determining the direction in the space corresponding to the abilities of interest, and finally estimating person abilities along those directions.

Our application of MIRT involves the Math and Reading Achievement sections of the Woodcock-Johnson Psycho-educational Battery-Revised (1989). Our discussion covers the test, the norming sample, our uni- and multidimensional calibration efforts, and ability estimation. Then we turn to the results comparing the theoretical estimates of information in the unidimensional and multidimensional applications.

Method

Subjects

The subjects came from the adults aged 20 - 39 in the

norming sample of the Woodcock-Johnson Psycho-educational Battery-Revised (Woodcock & Johnson, 1989). Because BILOG (Mislevy & Bock, 1989), the unidimensional item calibration program employed in this study, can handle no more than 1000 subjects, 1000 subjects aged 20 - 39 were randomly sampled from the full norming group in this age range.

Test Items

The Woodcock-Johnson Psycho-educational Battery-Revised (WJ-R, Woodcock & Johnson, 1989) contains individually administered, free response items designed to assess over a wide range of abilities. The WJ-R is a battery of aptitude and achievement tests that was normed on a nationally representative sample of 6,359 subjects from age 2 to 90+ years (McGrew, Werder & Woodcock, 1991). Items from three of the four WJ-R reading tests and two of the three mathematics tests were used in this study.' The very easiest ($p > .99$) and very hardest ($p < .01$) items were eliminated from the analysis. This left 36 reading items, composed of 19 Letter-word Identification and Word Attack items and 17 Passage Comprehension items, and 52 mathematics items, composed of 30 Calculation items and 22 Applied Problems.

Both the Calculation and Applied Problems subtests of the mathematics items start with easy items involving simple arithmetic calculation and advance to more difficult items involving algebra, trigonometry, logs, etc. The Letter-word Identification and Word Attack items start with easy items

involving identification of letters, then pronunciation of English words, and finally sounding out nonsense syllables according to conventions of English pronunciation; combined these items measure Basic Reading Skills that include both sight vocabulary and the ability to apply phonic and structural analysis skills. The Passage Comprehension subtest first involves finding a correct word to complete a simple sentence so the sentence accurately describe an accompanying picture. Harder items involve inferring the correct word to complete a short paragraph with no pictorial clues; subjects must exercise a variety of comprehension and vocabulary skills in this modified cloze testing procedure.

Calibration

BILOG (Mislevy & Bock, 1989) was used for the unidimensional item calibrations. Six unidimensional item sets were calibrated using BILOG: the 36 Reading Items, the 52 Mathematics items, the 19 Basic Reading Skills items, the 17 Passage Comprehension items, the 30 Calculation items, and the 22 Applied Problems.

One two-dimensional item calibration was conducted for all items using TESTMAP (McKinley, 1991a). Problems were encountered in obtaining a reasonable solution in more than two dimensions. Abilities were estimated using the companion program THETA (1991b). The initial estimates produced by TESTMAP were rotated so as to be orthogonal, and the ability estimates (θ_1, θ_2) along these orthogonal reference vectors were standardized to have mean

0 and variance 1.00. Estimates of item discrimination and difficulty parameters were rotated and rescaled so as to be consistent with the rotation and standardization of the ability estimates. The likelihood ratio fit statistic for TESTMAP was 60528.03.

For purposes of evaluating test information, thirty-four subjects were selected so as to cover the region with greatest density for a bivariate normal distribution centered at the origin. Figure 1 shows these 34 points.

Insert Figure 1 about here.

In the multidimensional space, directions corresponding to each content area were determined using the method of Wang (1986). For each individual and content area, a score was obtained by using the subject's reference vector scores (θ_1, θ_2) to estimate the subject's projection onto the vector corresponding to the particular content area. For each content area, the unit and origin of MIRT estimates were set according to the same convention employed by the unidimensional calibration program BILOG; that is, the origin was set to zero and the variance was set to 1.00.

For each of the 34 subject points, and each content area, the unidimensional test information was computed as the sum of the item information, where item information was computed

according to Equation 2 substituting sample estimates from the item calibration for population parameters. Multidimensional information was computed as the sum of item information, where item information is computed as in Equation 4, again substituting sample estimates from the MIRT calibration for the population parameters. For both UIRT and MIRT, the standard error at a given ability θ was computed from test information, Equation 5.

Results

Figure 2a shows the MIRT discrimination parameters broken down so as to distinguish the Reading and Math items. Reading items generally have higher discriminations along Dimension 1 than along Dimension 2, and hence we shall call this the Verbal Reference Vector, but it should be understood that it does not exactly coincide with the vector best representing the Reading content area. We will call Dimension 2 the Math Reference Vector, although most math items have roughly equal discrimination parameters on the two dimensions.

Figure 2b shows the discrimination parameters broken down so as to distinguish Basic Reading Skills and Passage Comprehension items. The two verbal measures, Passage Comprehension and Basic Reading Skills lie in approximately the same directions in the space. Figure 2c shows the discrimination parameters for the Calculation and Applied Problems mathematics items. These two sets of items are visually distinct, although they lie along

similar vectors in the space. This means that as compared to the Calculation items, the Applied Problems have higher discrimination parameters on the Verbal Reference Vector and lower discrimination parameters on the Math Reference Vector. Generally, the pattern of discrimination parameters did not sharply distinguish the subtests within Reading and within Math.

Insert Figures 2a - 2c about here.

Table 1 shows the intercorrelations of UIRT and MIRT estimates of achievement in the six content areas. The intercorrelations of the UIRT and MIRT estimates, the diagonal of Table 1, are not as high as one might expect, given that they are scores for the same content area and given that the same item responses which determine the UIRT score also comprise part of the items determining the MIRT score. Correlations range from .83 for the Basic Reading Skills and Passage Comprehension to .94 for Math.

Within the UIRT and MIRT estimates, the patterns of correlations are noticeably different. That is, the intercorrelations of UIRT estimates (below the diagonal) are substantially lower than the intercorrelations of MIRT estimates (above the diagonal). The MIRT reading, Basic Reading Skills and Passage Comprehension scores are virtually 1.00, as are the

intercorrelations of Math, Calculation, and Applied Problems. Since the pattern of item discriminations (Figures 2b and 2c) were so similar for subareas within Reading and Math, the multidimensional ability estimates for the content areas within Reading and Math were virtually identical.

Insert Table 1 about here.

Figures 3a, 3b, and 3e show unidimensional and multidimensional standard errors as a function of unidimensional scores for Reading and its two subareas, Basic Reading Skills and Passage Comprehension. Each point corresponds to one of the 34 subjects in Figure 1. Multidimensional standard errors are generally smaller than corresponding unidimensional standard errors. This difference between unidimensional and multidimensional standard errors is larger for the Basic Reading Skills and Passage Comprehension subareas than for the Reading area as a whole.

In the Mathematics areas; Figures 3c, 3d, and 3f; the MIRT standard errors are usually, but not always, smaller. At high levels of ability, the UIRT standard errors are actually smaller than the MIRT standard errors. This led us to inspect more closely the MIRT and UIRT scores in corresponding content areas.

Insert Figures 3a - 3f about here.

To prepare Figures 4a and 4b, the 1000 subjects were blocked into 20 score groups with 50 in each group based on their UIRT scores in Reading and Math. For each score group, it's means on reference vectors θ_1 and θ_2 were plotted in either Figure 4a or 4b. If the UIRT scores corresponded to vectors in the MIRT space, each of these plots would form a straight line. The UIRT Reading scores, Figure 4a, do fall roughly (and arguably) along a straight line. However, the Math scores clearly form a curvilinear surface. The scores appear to heavily covary with the Verbal Reference Vector at low ability levels, but appear to be orthogonal to the Verbal Reference Vector at higher ability levels. The MIRT Math scores, by definition, are projections along a straight line in this space; the UIRT Math scores are not. Thus the MIRT and UIRT Math scores index linear and nonlinear composites respectively of the MIRT reference vectors. This helps explain why the correlation between MIRT and UIRT Math scores in Table 1 is lower than one might expect for scores in a common content area.

Insert Figures 4a - 4b about here.

Conclusions

Unidimensional IRT theories can always be represented as hierarchically embedded in a multidimensional compensatory model. In this multidimensional representation, the unidimensional ability is linearly related to the multidimensional reference vectors. As our Math items illustrate, however, empirical unidimensional ability estimates need not be linearly related to reference vectors in an empirically derived multidimensional representation of the full test battery. Thus, one can reasonably expect multidimensional scores to be consistently more precise estimates of the construct indexed by unidimensional scores only when the unidimensional scores are linearly related to the multidimensional reference vectors as the compensatory model would lead one to expect. In any application of MIRT to capitalize on collateral information, the researcher needs to examine the relationship between unidimensional scores and multidimensional reference vectors. Unidimensional discrimination parameters should be linearly related to multidimensional discrimination parameters, and unidimensional abilities should be a linear function of multidimensional abilities.

The MIRT approach will lead to more accurate estimates of subtest scores in a larger test, only when there are distinctly different patterns of item discriminations for the various subtests. Within Math and Reading, items did not differ by

subtest in their discrimination patterns. Therefore, the multidimensional approach did not yield scores within Math and Reading which were distinct from the Math and Reading scores themselves. In any application of MIRT to capitalize on collateral information, the researcher needs to examine the patterns of discrimination parameters for subareas within a broader item set to determine if the discrimination patterns are distinctly different.

In short, for the MIRT approach based on a compensatory model to effectively capitalize on collateral information, empirically derived scores must satisfy the following criterion: scores in each subarea need to be representable as distinct vectors in the multidimensional representation of items. Further research is needed to determine which types of items, content areas, and subareas conform to this criterion.

In our analysis, unidimensional Reading scores were (arguably) representable as a linear function of the multidimensional reference vectors. Multidimensional standard errors were generally less than or equal to the conditional standard errors for the unidimensional estimates. Hence the multidimensional Reading scores can reasonably be considered more precise estimates of the ability indexed by the unidimensional Reading scores. However, reading subareas, Basic Reading Skills and Passage Comprehension were not represented by distinctly different patterns of item discriminations. Hence our

multidimensional representation yielded more precise scores, but not scores identifiable as Basic Reading Skills and Passage Comprehension distinct from Reading.

Math items may consistently violate the above criterion. In our analysis, the unidimensional math scores were nonlinearly related to the multidimensional reference vectors in our two-dimensional representation. Ackerman (1991) alludes to similar anomalies in American College Test math data. Atkin, Bray, Davison, Herzberger, Humphreys, and Selzer (1977) found evidence for a factor differentiation hypothesis in which math and verbal factors seem to become less highly correlated with age. It may be, however, that the more difficult tasks commonly included in mathematics tests for older ages tap a different ability composite than do the easier tasks used to tap math ability at younger ages. This would explain the evidence both for factor differentiation reported by Atkin et al. and the changing composition of UIRT math scores with increasing ability reported in the present study. If arithmetic items and mathematics items (formal algebra, trigonometry, etc.) tap different factors, then unidimensional mathematics scores may not always be representable as simple linear composites of empirically derived multidimensional ability vectors, and hence mathematics scores may not capitalize on collateral information in the fashion predicted by compensatory MIRT models.

References

Ackerman, Terry A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. Applied Psychological Measurement, 15, 13 - 24.

Ackerman, T. A. & Davey, T. C. (1991). Concurrent adaptive measurement of multiple abilities. Paper presented to the National Council on Measurement in Education, Chicago.

Atkin, R., Bray, R., Davison, M., Herzberger, S., Humphreys, L., & Selzer, U. (1977). Ability factor differentiation, grades 5 through 11. Applied Psychological Measurement, 1, 65 - 76.

Chang, Y.W. & Davison, M. L. (1991). On test information: A comparison of the multidimensional and unidimensional item response theory. Proceedings of the International Educational Statistics and Measurement Symposium. Tainan, Taiwan, R.O.C.: Tainan Teachers College.

Davey, T. & Hirsch, T. M. (1991). Concurrent and consecutive estimates of examinee ability profiles. Paper presented to the annual meeting of the Psychometric Society, New Brunswick, NJ.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233 - 245.

Luecht, R. M. & Miller, T. R. (in press). Unidimensional calibrations and interpretations of composite abilities for multidimensional tests. Applied Psychological Measurement.

McGrew, K. S., Werder, J. K. & Woodcock, R. W. (1991). Woodcock-Johnson Psycho-educational Battery-Revised Technical Manual. Allen, TX: DLM.

McKinley, R. C. (1991a). TestMAP Version 2.0 User's Guide. Princeton, NJ: Educational Testing Service.

McKinley, R. C. (1991b). User's Guide to THETA PC Version 2.0. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Bock, R. D. (1989). PC-BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models. Mooresville, IN: Scientific Software.

Reckase, M. D. (1985). The difficulty of test items that measure more than one dimension. Applied Psychological Measurement, 9, 401 - 412.

Reckase, M. D. (1986). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293 - 325.

Woodcock, R. W. & Johnson, M. B. (1989). Woodcock-Johnson Psycho-educational Battery-Revised. Allen TX: DLM.

Wang, M. M. (1986). Fitting a unidimensional model to multidimensional response data. Paper presented at the annual Office of Naval Research Contractor's Meeting. Knoxville, TN.

Table 1. Intercorrelations of Unidimensional and Multidimensional Ability Estimates

	BRS	PC	C	AP	R	M
BRS	.83	1.00	.68	.82	1.00	.73
PC	.65	.83	.63	.78	1.00	.69
C	.44	.48	.90	.98	.64	1.00
AP	.51	.59	.74	.87	.78	1.00
R	.92	.88	.50	.60	.92	.69
M	.51	.56	.95	.90	.59	.94

Note: Correlations of corresponding multi- and unidimensional ability estimates are given on the diagonal. Intercorrelations of multidimensional estimates are shown above the diagonal. Intercorrelations of unidimensional estimates are shown below the diagonal. BRS = Basic Reading Skills, PC = Passage Comprehension, C = Calculation, AP = Applied Problems, R = Reading, and M = Math.

Figure Captions

Figure 1. Locations of the 34 points used to compare unidimensional and multidimensional information.

Figure 2a. Multidimensional discrimination parameters for the Reading and Math Test items; R = Reading Item, M = Math Item.

Figure 2b. Multidimensional discrimination parameters for the Reading Test items; B = Basic Reading Skills Item, P = Passage Comprehension Item.

Figure 2c. Multidimensional discrimination parameters for the Mathematics Test items; C = Calculation Item, A = Applied Problems Item

Figure 3a. Standard errors for unidimensional and multidimensional Basic Reading Skills scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

Figure 3b. Standard errors for unidimensional and multidimensional Passage Comprehension scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

Figure 3c. Standard errors for unidimensional and multidimensional Calculation scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

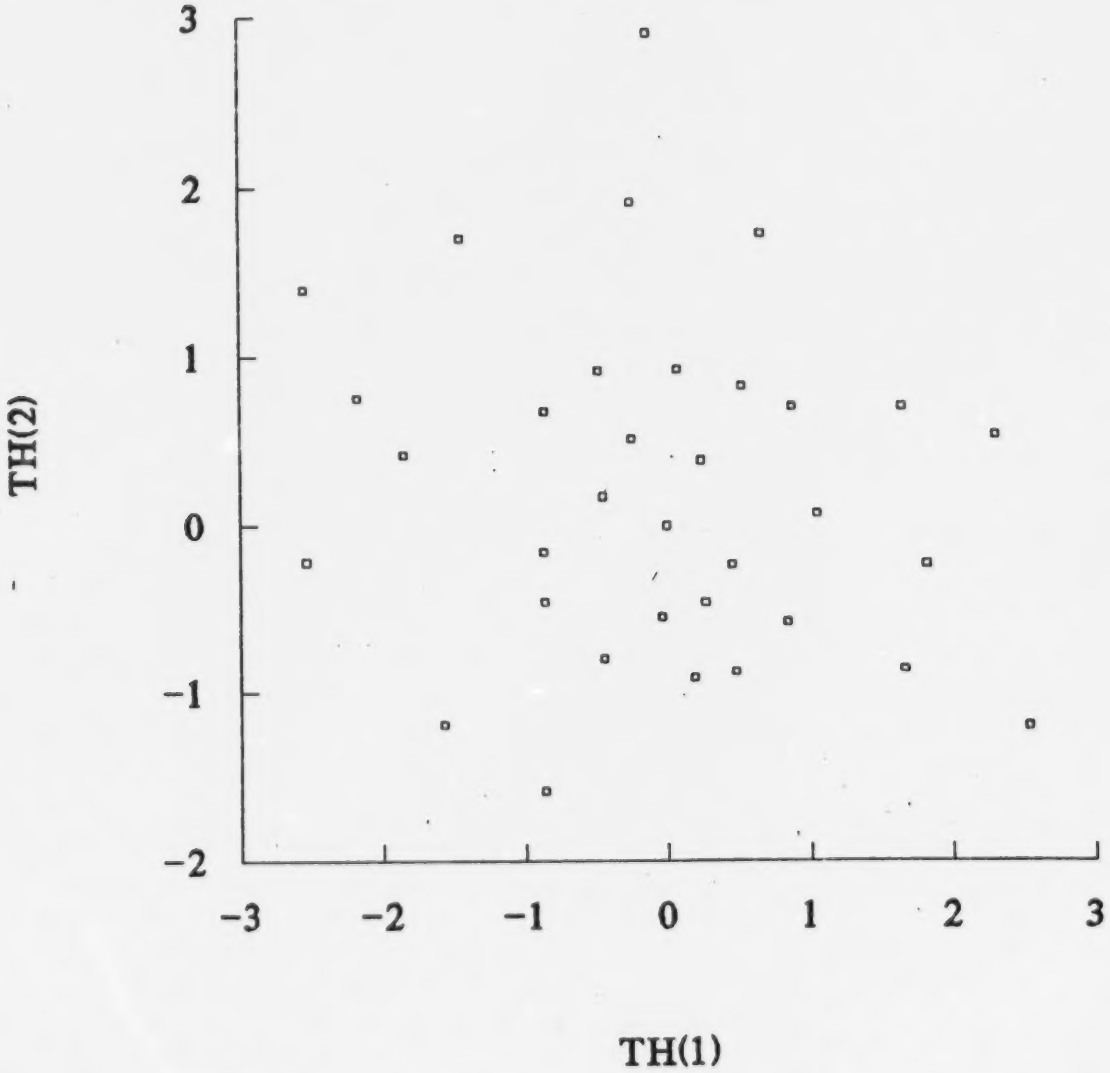
Figure 3d. Standard errors for unidimensional and multidimensional Applied Problems scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

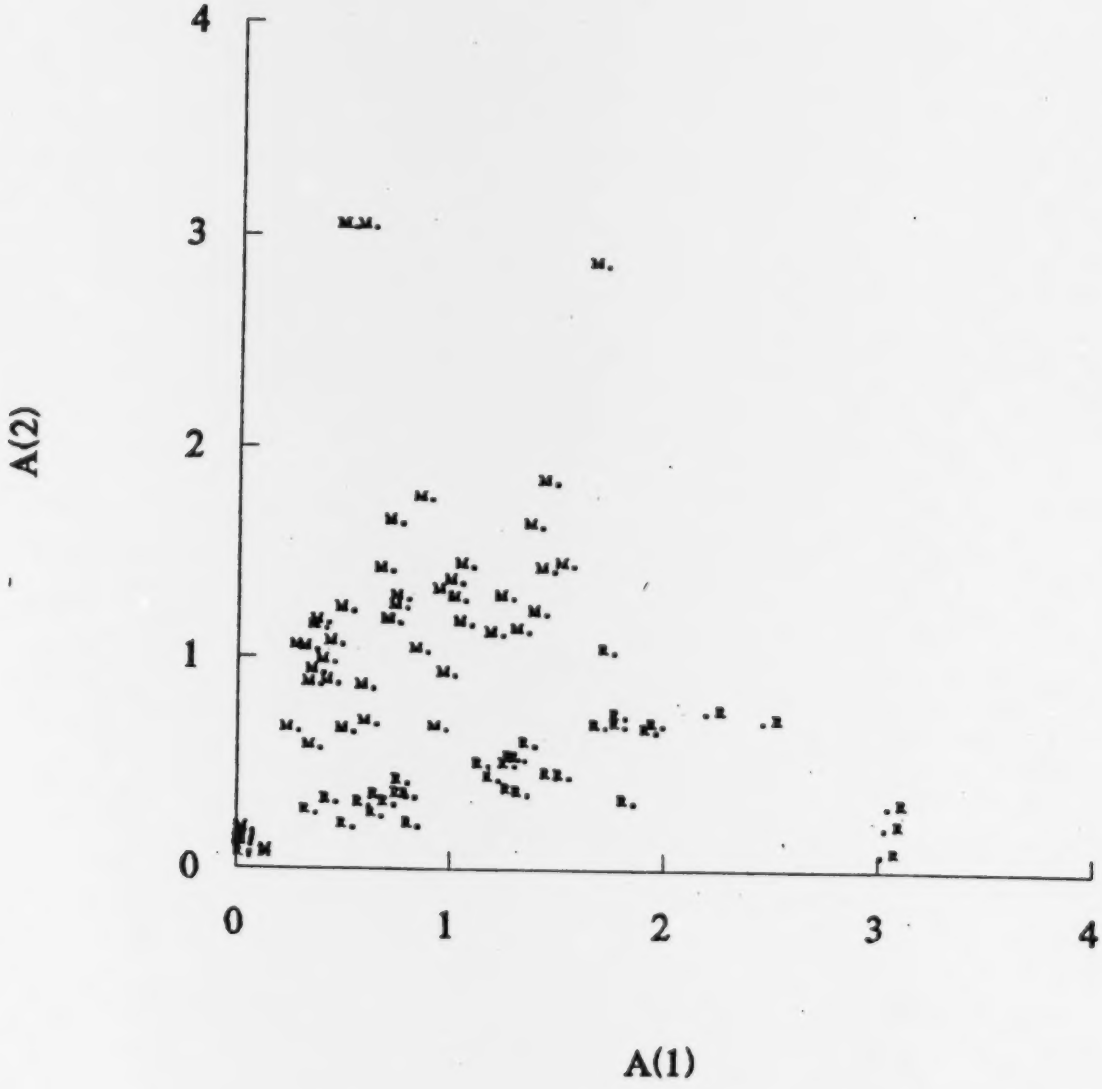
Figure 3e. Standard errors for unidimensional and multidimensional Reading scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

Figure 3f. Standard errors for unidimensional and multidimensional Mathematics scores plotted against unidimensional scores for the 34 subjects in Figure 1. U = Unidimensional and M = Multidimensional.

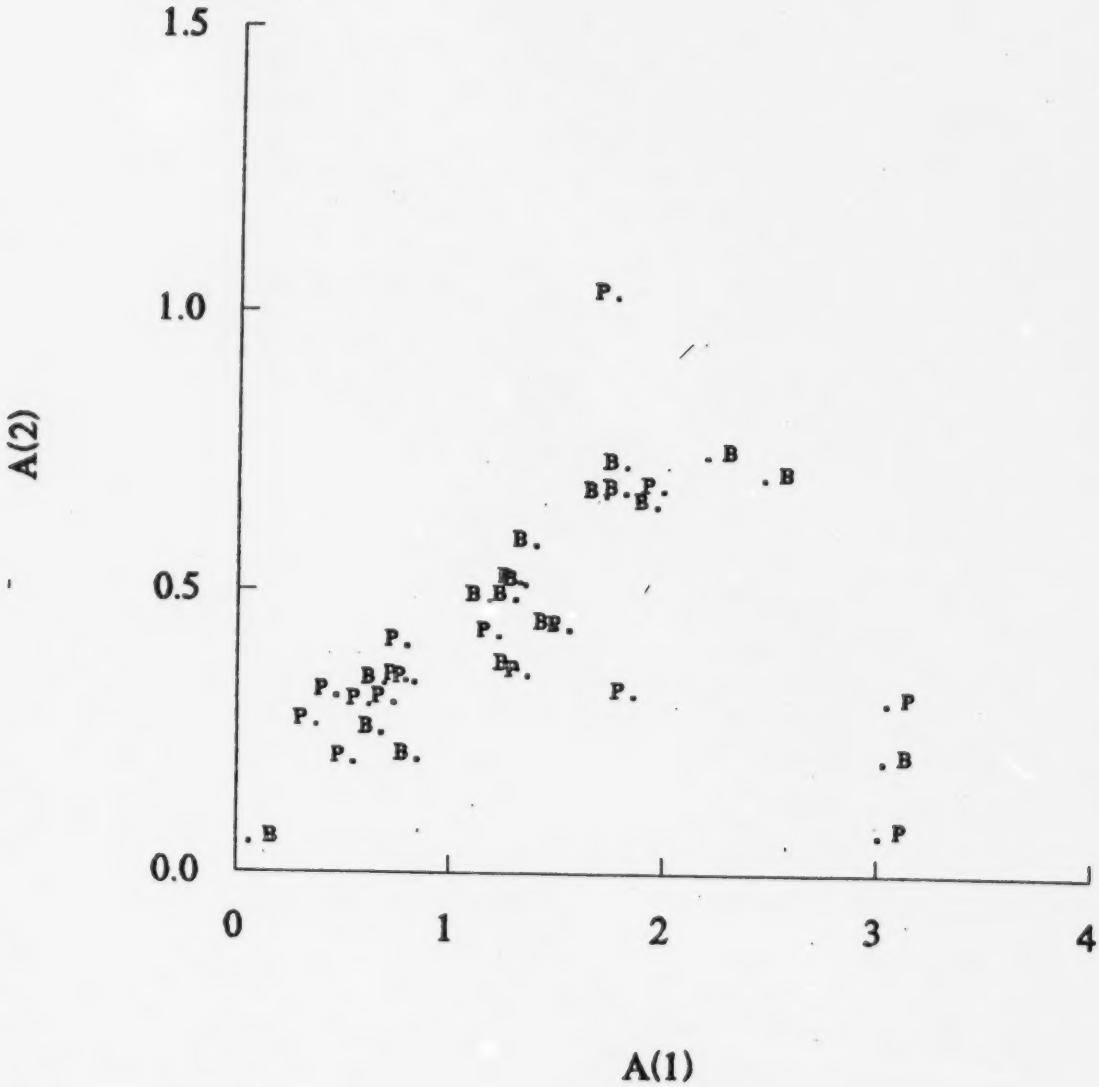
Figure 4a. Mean scores on reference vectors θ_1 and θ_2 for subjects blocked by unidimensional Reading scores.

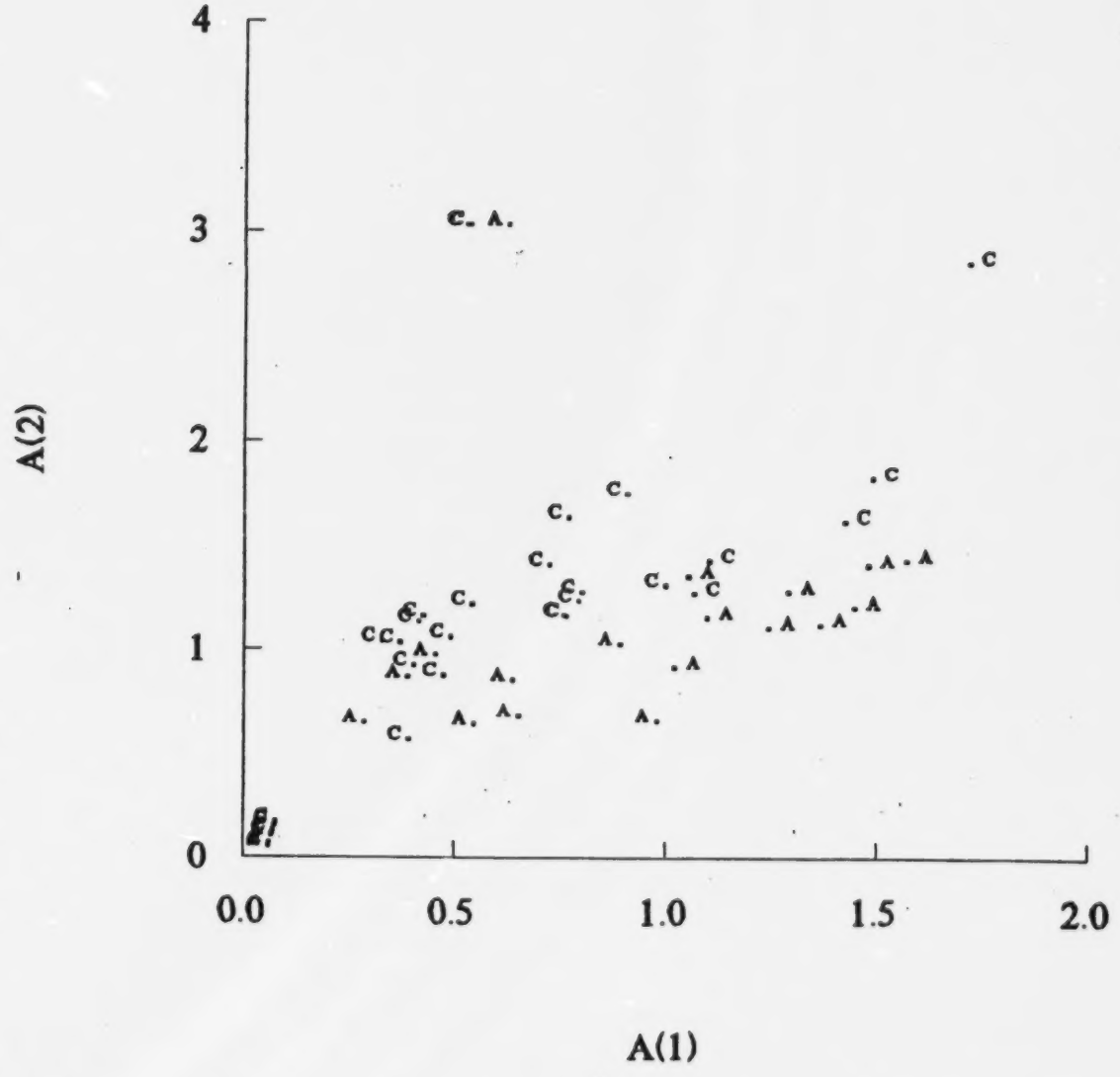
Figure 4b. Mean scores on reference vectors θ_1 and θ_2 for subjects blocked by unidimensional Math scores.

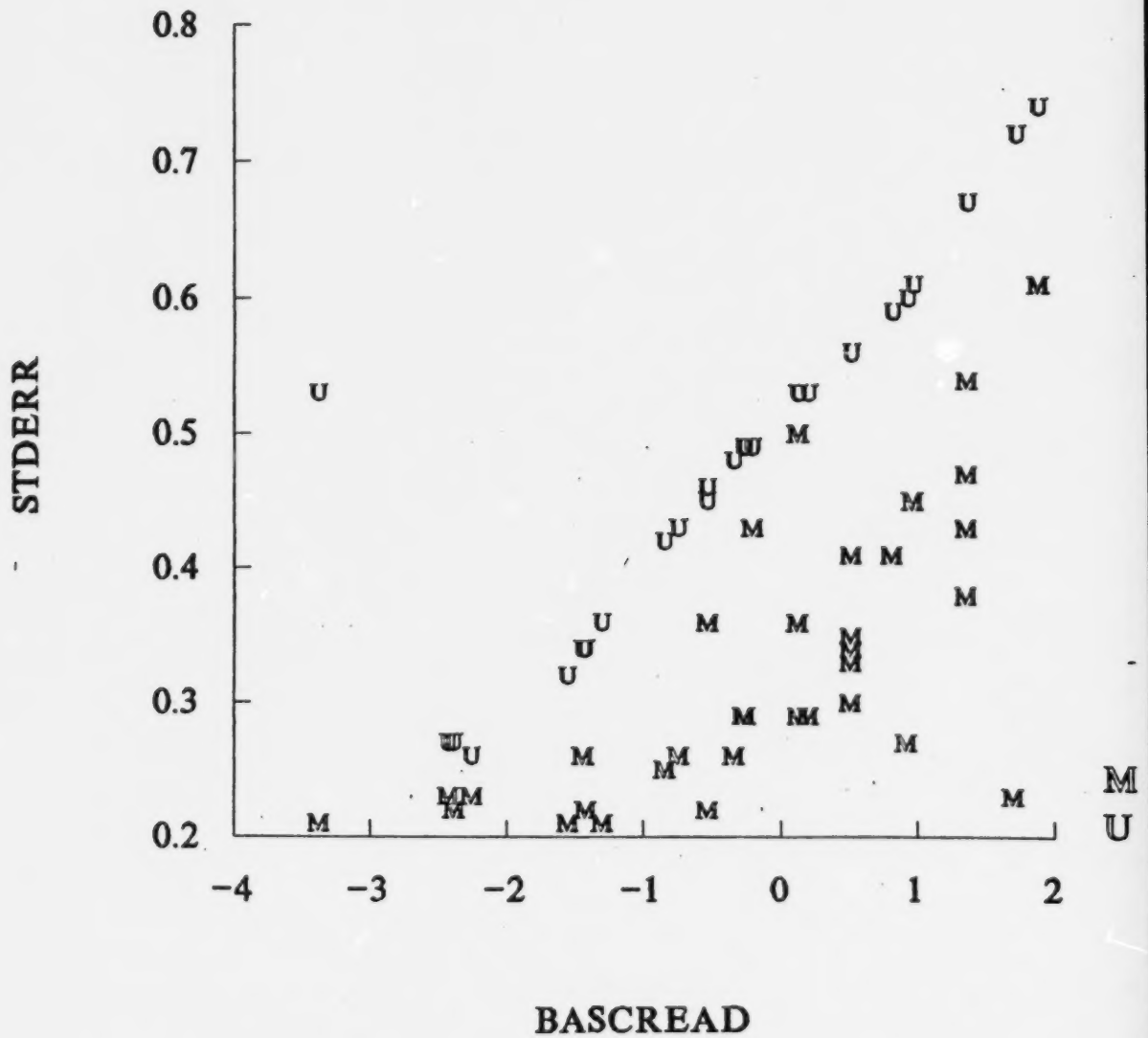

3
2
1
0
−1
−2
TH(2)
−3
−2
−1
0
1
2
3
TH(1)

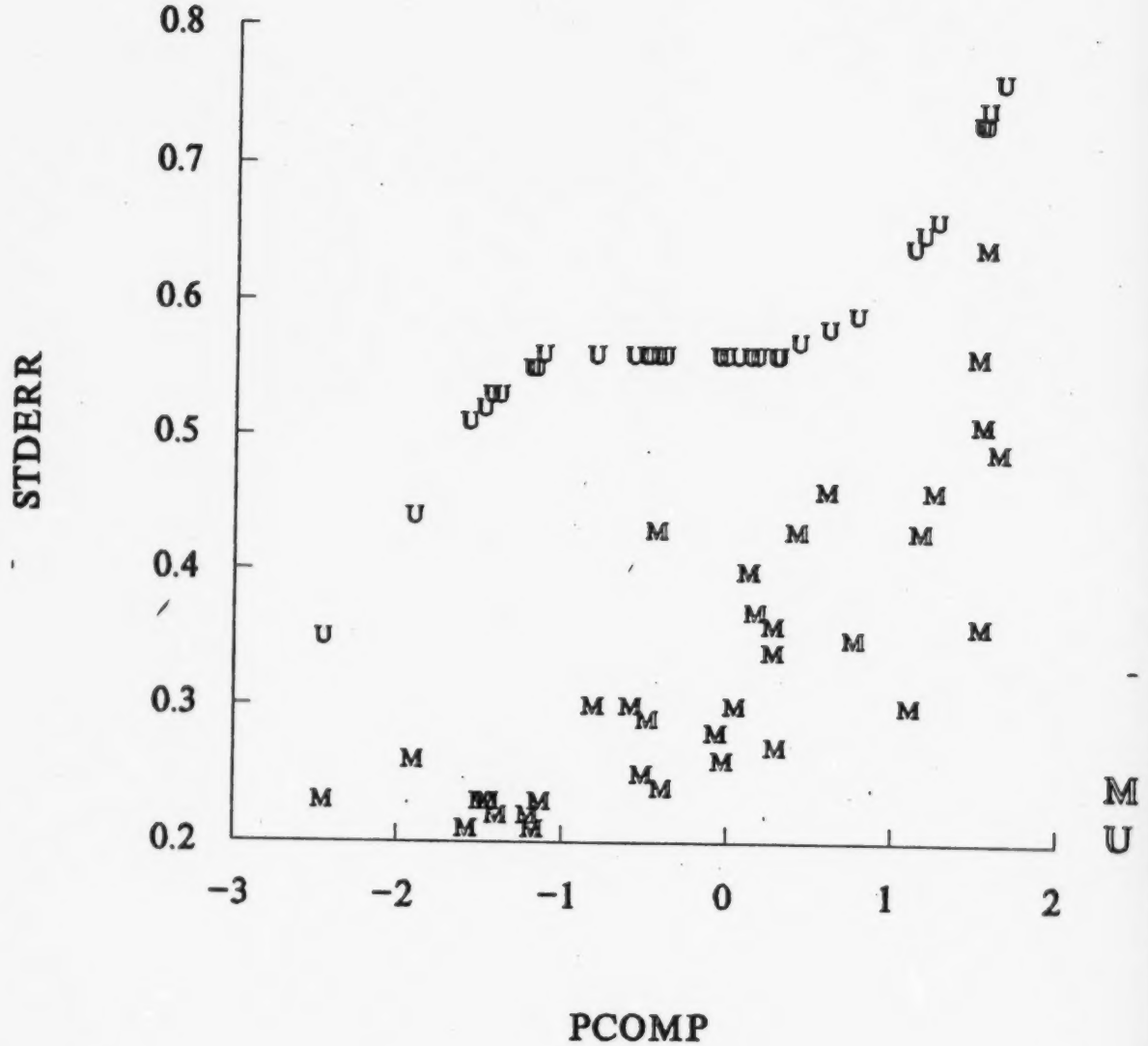


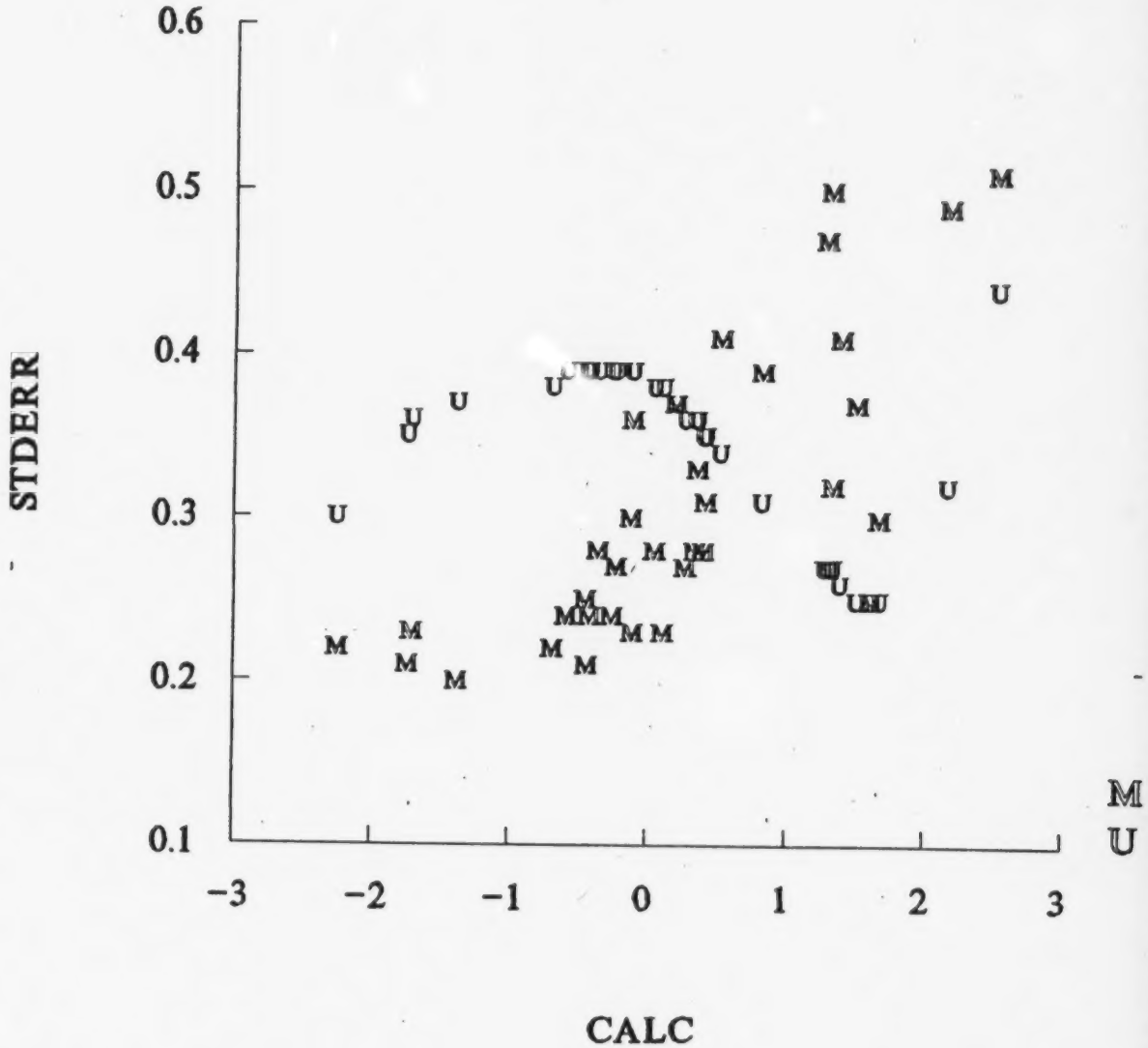
BEST COPY AVAILABLE

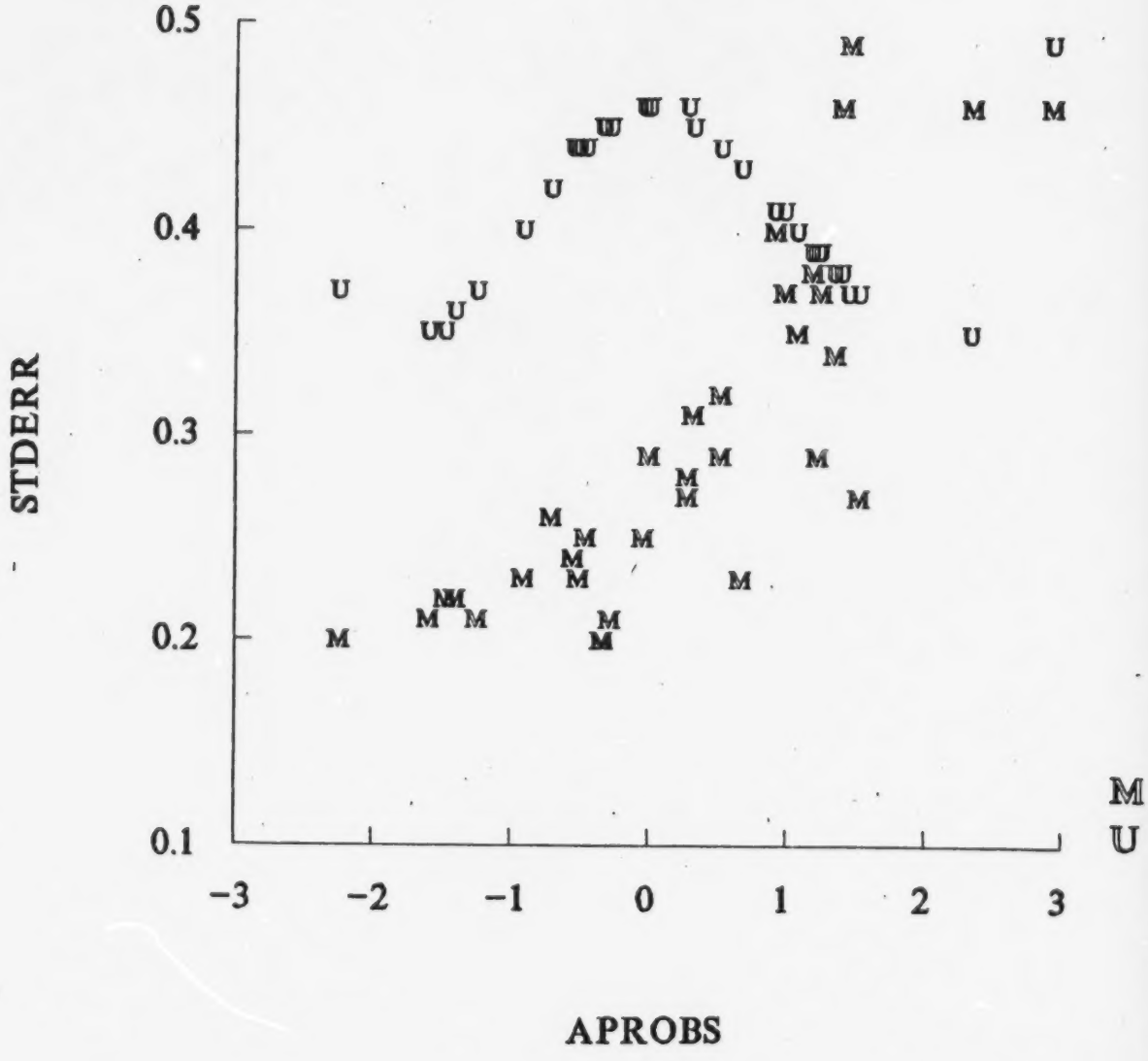

1.5
1.0
0.5
0.0
A(2)
0
1
2
3
4
A(1)

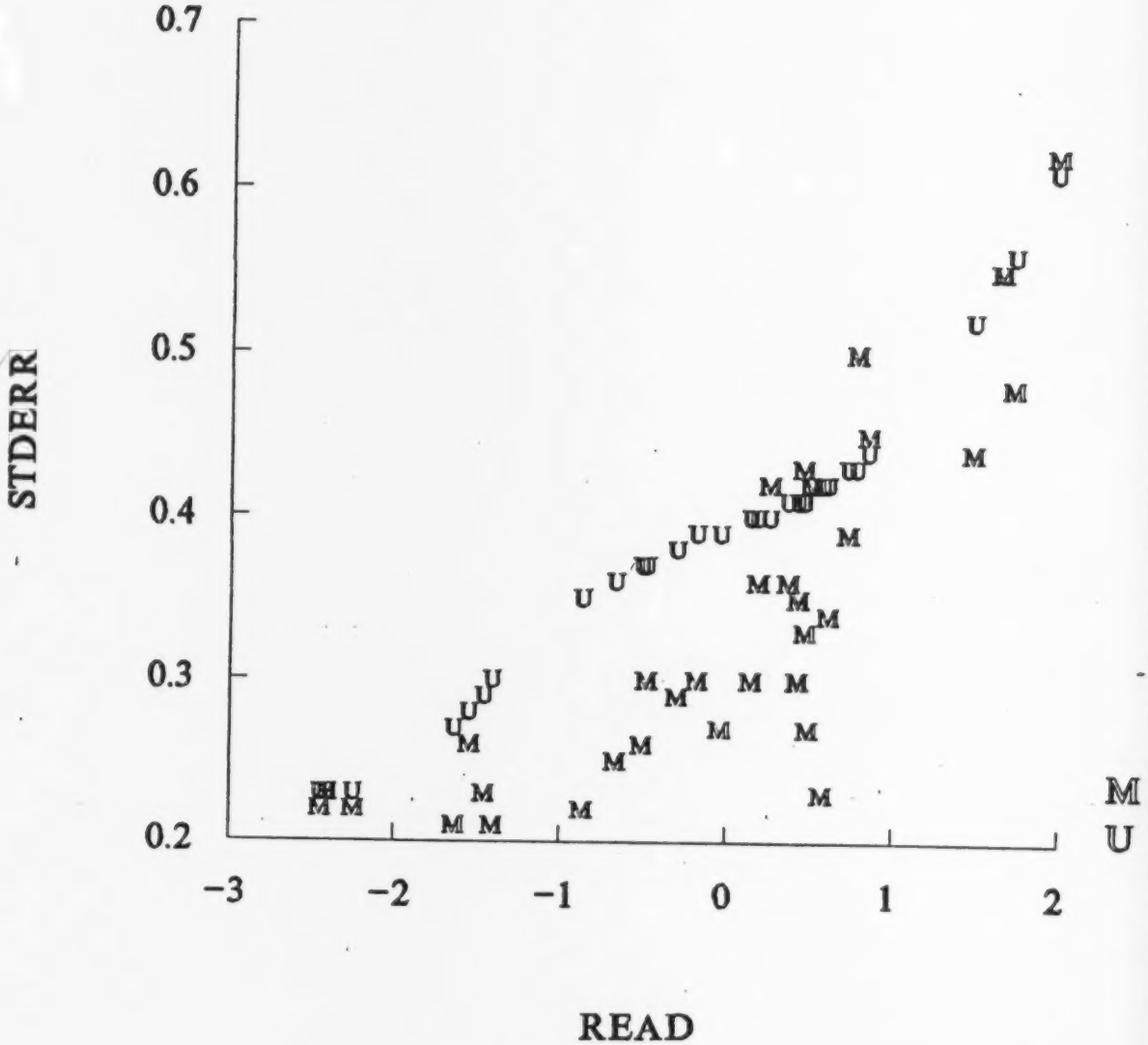

4
3
2
1
0
A(2)
0.0
0.5
1.0
1.5
2.0
A(1)

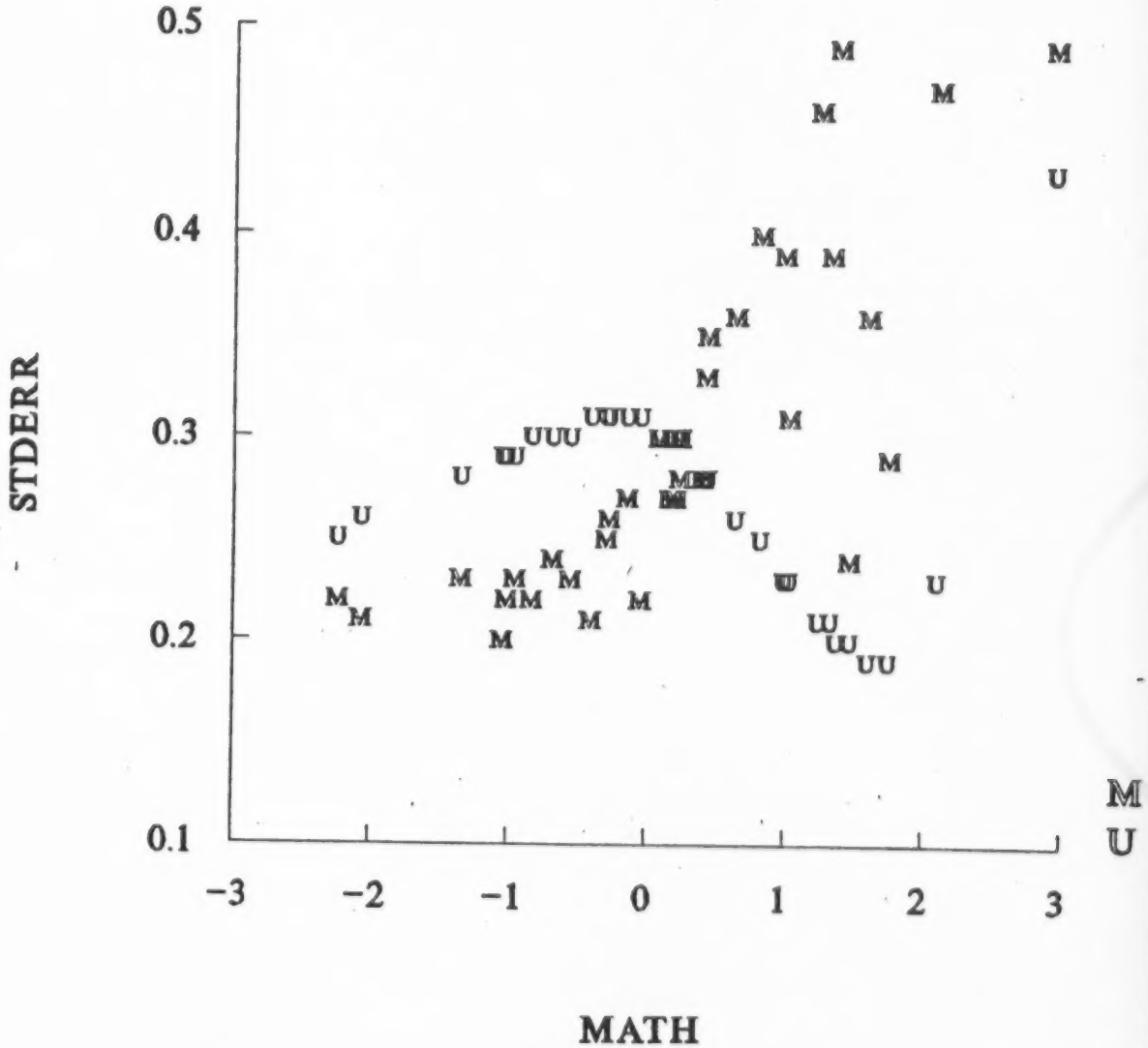

STDERR
0.8
0.7
0.6
0.5
0.4
0.3
0.2
−4
−3
−2
−1
0
1
2
BASCREAD

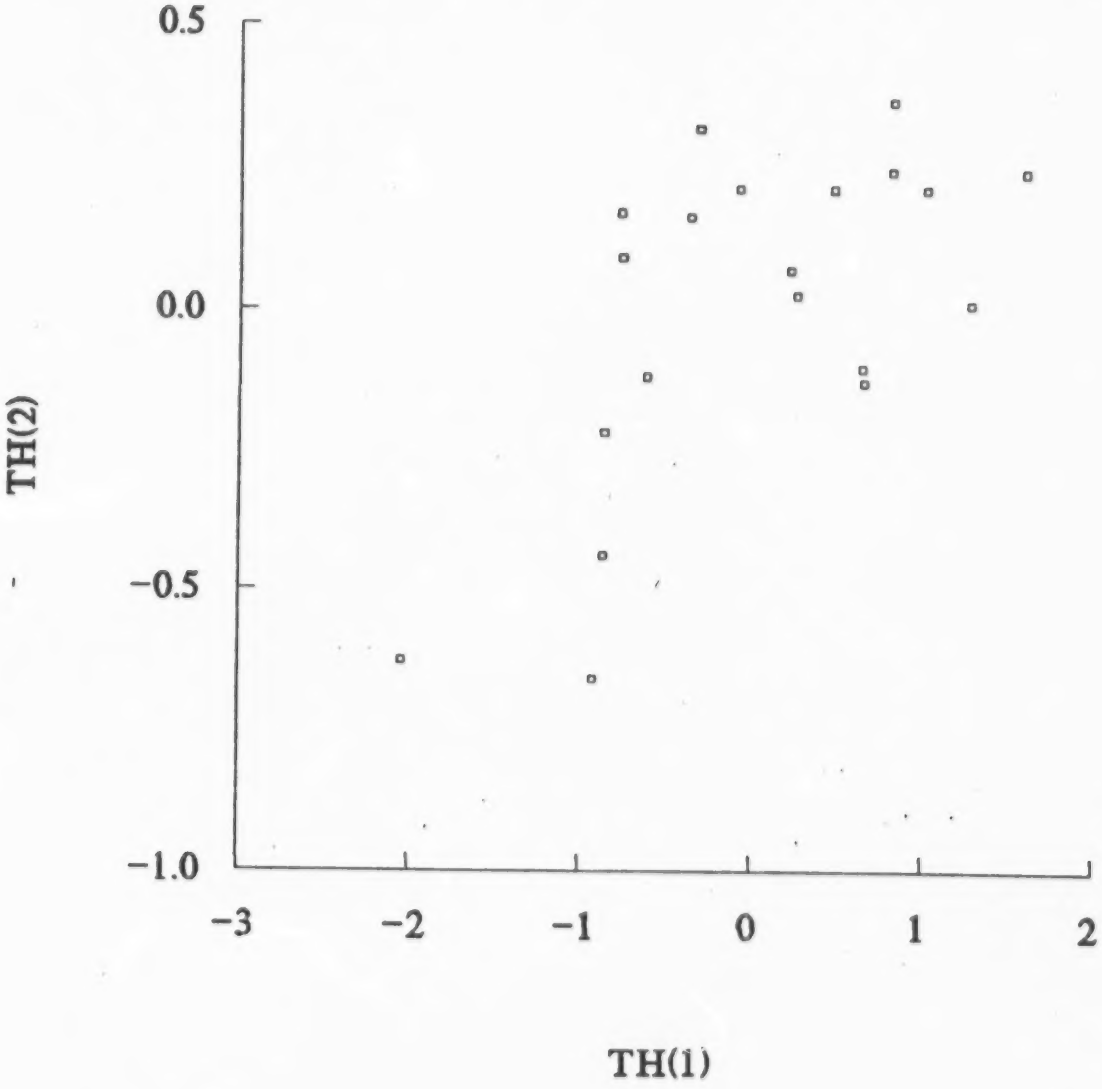

STDERR
0.8
0.7
0.6
0.5
0.4
0.3
0.2
−3
−2
−1
0
1
2
PCOMP
M
U

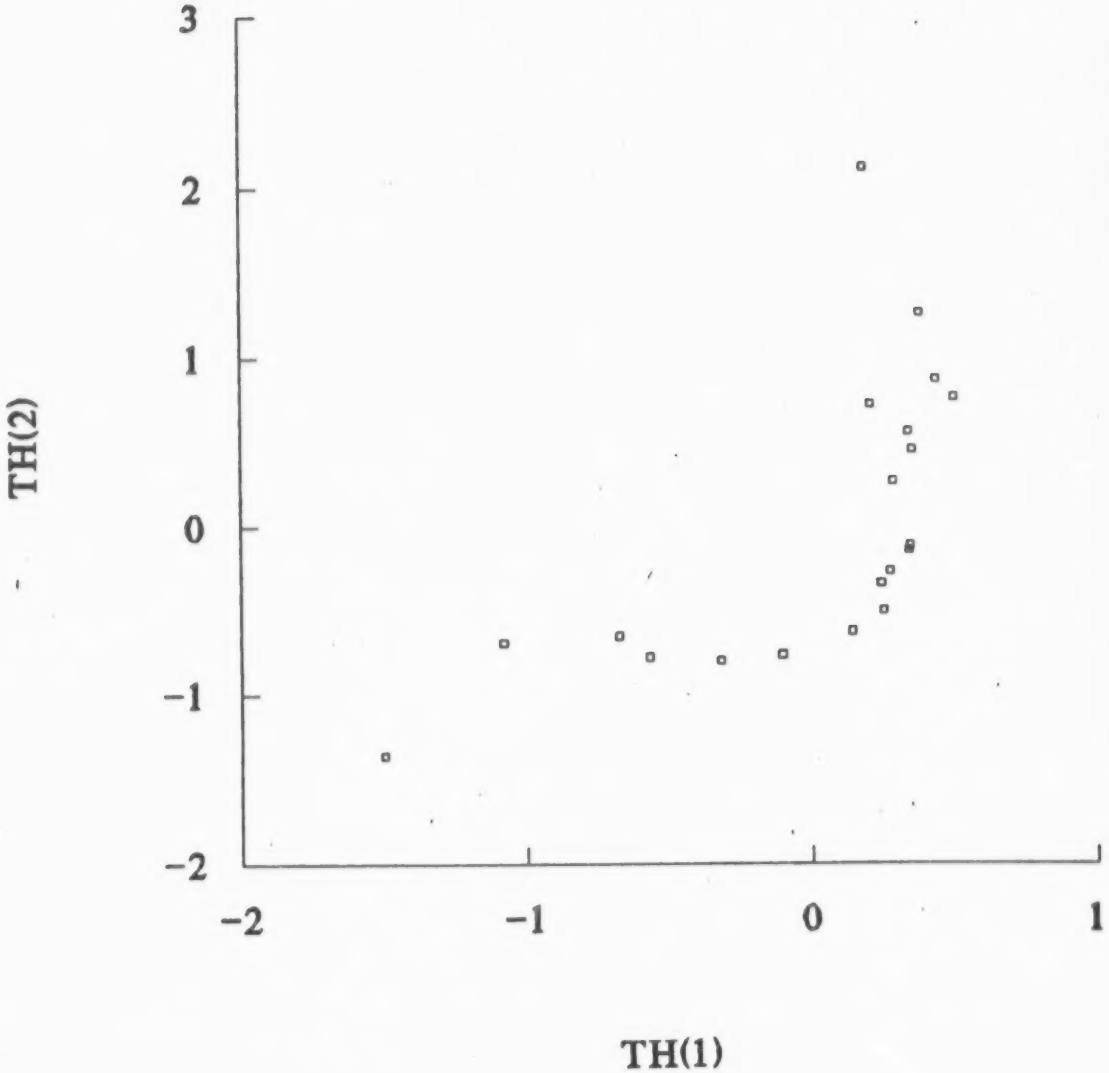

0.6
0.5
0.4
0.3
0.2
0.1
STDERR
−3
−2
−1
0
1
2
3
CALC
M
U


0.5
0.4
0.3
0.2
0.1
STDERR
−3
−2
−1
0
1
2
3
APROBS
M
U


STDERR
0.7
0.6
0.5
0.4
0.3
0.2
−3
−2
−1
0
1
2
READ


0.5
0.4
0.3
0.2
0.1
STDERR
−3
−2
−1
0
1
2
3
MATH
M
U


0.5
0.0
−0.5
−1.0
TH(2)
−3
−2
−1
0
1
2
TH(1)


3
2
1
0
−1
−2
TH(2)
−2
−1
0
1
TH(1)